Performance Analysis and Design Validation

A Synthesized Evaluation of Agentic Proof Strategies

Dietmar Wolz, Ingo Althöfer

February 14, 2026

Abstract

This document synthesizes the results of our autonomous AI-assisted proof attempts for the "First Proof" challenge (Questions 1–10). By comparing our generated solutions [2] against the official solutions and conducting a deep technical audit, we categorize the performance into three distinct tiers: **Logical Failures** (semantic misunderstandings), **Heuristic Traps** (analytical shortcuts), and **Algorithmic Wins** (superhuman optimization). These findings empirically validate the necessity of the Human-in-the-Loop (HITL) architecture proposed in our design document [1].

1 Category I: Logical and Semantic Failures

*These failures stem from the agent's inability to parse subtle logical quantifiers or definitions. They highlight the necessity of the **Human Manager** role.*

Question 2: Generic Representations (The Quantifier Trap)

Status: Failed

- **Divergence:** The problem required a single Whittaker function W effective for *all* representations ($\exists W, \forall \pi$). The agent constructed a W dependent on π ($\forall \pi, \exists W$).
- **Root Cause:** The agent optimized for the "easier" interpretation of the predicate.
- **Design Validation:** A human manager is required to perform **Structured Decomposition**, explicitly locking the quantifier order before generation begins.

Question 7: Acyclic Universal Covers (The Definition Gap)

Status: Failed

- **Divergence:** The agent assumed "acyclic" implied "integrally acyclic" (over $\mathbb{Z}$), leading to a proof that no torsion exists. The official problem allowed "rationally acyclic" (over $\mathbb{Q}$), which permits torsion (e.g., Fowler's manifolds).
- **Root Cause:** Semantic ambiguity. The agent hallucinated a stricter constraint to enable the use of standard Smith Theory tools.
- **Design Validation:** The **Manager** must act as a semantic anchor, resolving definitions (Rational vs. Integral) against the problem statement.

Question 8: Lagrangian Smoothing (The Gluing Error)

Status: Failed

- **Divergence:** The agent attempted to glue local solutions by simply summing Hamiltonians ($H = \sum H_i$). The official solution required a "conormal fibration" to handle the geometric obstruction.
- **Root Cause:** "Linearity Bias." The agent assumed local charts could be combined linearly, ignoring the manifold structure.

2 Category II: Heuristic Traps and Hallucinations

*These solutions were structurally plausible but failed in rigorous details, often hallucinating machinery or citing sketches as proofs. They highlight the need for the **Reference Checker** and **Verifier**.*

Question 1: Mutual Singularity of Φ^4_3

Status: Partial / Rigor Deficient

- **Technical Nuance:** The agent used a **super-exponential scale** ($\varepsilon_n = \exp(-e^n)$) to force convergence heuristically. The official solution required **logarithmic scales** and delicate Borel-Cantelli estimates.
- **The Citation Trap:** The agent cited an unpublished note by Hairer to skip the hardest step (stability of the measure).
- **Design Validation:** The **Reference Checker** must verify that cited lemmas are rigorous theorems, not sketches. The agent optimized for "plausibility" over "analytical truth."

Question 4: Finite Free Stam Inequality

Status: Hallucinated Machinery

- **Technical Nuance:** The agent attempted to use a "conditional expectation" argument analogous to the classical case. It missed the specific tool required: the geometry of **Hyperbolic Polynomials** and real-rootedness preservation.
- **Root Cause:** False Analogy. The agent assumed a classical probability tool had a direct non-commutative counterpart.

Question 3: TASEP Stationary Distribution

Status: Correct but Trivial

- **Divergence:** The agent used Metropolis-Hastings to reverse-engineer a sampler. The problem sought the "Push TASEP" interacting particle system.
- **Design Validation:** Agents lack "mathematical taste." A human must guide the **Strategy Selection** to ensure the solution reveals structural insight, not just numerical correctness.

3 Category III: Algorithmic Wins (Superhuman Results)

*In computational tasks, the autonomous agent matched or outperformed human experts. This validates the **Generator's** role in algorithmic search.*

Question 10: Matrix-Free Tensor Completion

Status: Superior to Human Solution

- **The Win:** The agent proposed a **Matrix-Free** approach with $O(qnr)$ complexity, which the problem author (Kolda) explicitly preferred over her own $O(n^3)$ solution.
- **The Nuance:** While the solver was superior, the agent's choice of a Kronecker preconditioner was computationally heavier than the reference's simple diagonal preconditioner.
- **Design Validation:** The agent excels at searching the space of numerical algorithms. The HITL system should unleash the agent on optimization sub-problems while the human critiques the specific components (like the preconditioner).

Question 9: Identifiability of Rank-1 Tensors

Status: Correct and Optimal

- **The Win:** The agent correctly identified the 5×5 flattening minors and provided a rigorous Lie Algebra proof.
- **Validation:** Labeled "Best AI-generated solution" by the authors.

4 Design Implications: The Symbiotic Architecture

The empirical results from this challenge strictly map to the components of our design document [1]:

1. **The Manager (Human):** Essential for **Q2, Q7, Q8**. The agent cannot reliably distinguish between "Rational" and "Integral" acyclicity or handle complex quantifier nesting without human semantic anchoring.
2. **The Verifier (Human/Formal):** Essential for **Q1, Q4, Q6**. The agent will use heuristics (super-exponential scales) or hallucinations (non-commutative conditional expectation) to close gaps. A rigid verification step is required to catch these "plausible" errors.
3. **The Generator (Agent):** Validated by **Q9, Q10**. When the problem is well-defined and algorithmic, the agent acts as a "super-calculator," finding optimizations ($O(qnr)$ matrix-free methods) that human experts miss.

Conclusion: The future of automated reasoning is not fully autonomous. It is a hybrid system where humans provide the *Definition* and *Taste*, and agents provide the *Scale* and *Computation*.

References

[1] Wolz, D. (2026). *Agentic Strategy Design for Math Proofs*. Available at `https://althofer.de/agentic_strategy_design_for_math_proofs.pdf`. Accessed February 12, 2026.

[2] Wolz, D. (2026). "Multi-Agent Strategy for Solving Research-Level Mathematics." `https://althofer.de/first-proof-competition/first-proof-report.html`